CS 340: Computer Systems

Muhammed Can Özdemir
Ahsen Akpınar
Ece Uslu
Nejat Günaydın

Progress Presentation

Dataset details

- «Sınıflama» (Classification)
- « Eser Adı» (Book Name)
- «Yazar» (Author)
- «Dil» (Language)
- «Konu Başlıkları» (Title of subject)
- «Ödünç Sayısı» (Number of borrowing)

Dataset details

İSTANBUL ŞEHİR ÜNİVERSİTY

- Schema of the Data

```
df.printSchema()
```

```
root
 |-- Dil: string (nullable = true)
 |-- Eser Adı: string (nullable = true)
 |-- Konu Başlıkları: string (nullable = true)
 |-- Sınıflama: string (nullable = true)
 |-- Yazar: string (nullable = true)
 |-- Ödünç Sayısı: string (nullable = true)
```

What: the problem of interest

İSTANBUL ŞEHİR ÜNİVERSİTY

Our main interest is to build a program that does **similarity analysis** for books in the library of İstanbul Şehir University, so that can be used in a more complicated recommendation algorithm.

Why: reason, importance

İSTANBUL ŞEHİR ÜNİVERSİTY

The system that our university's library uses doesn't permit to get details about searchs and access info to each books (with or without user info). As a result, in order to access to more details, there is a need to build (or use) a more data-accessible system that our university's library has a total control on it.

The reason why we decided to make a program that make **similarity analysis** between books is to make a starter point to make our own recommendation system so that we have a total control on it.

How: the methodology

İSTANBUL ŞEHİR ÜNİVERSİTY

- What systematic approaches are you going to employ?

- First of all, we took our data as an excel file. After that we used xlrd module and each cell read by manually one by one. After that, it converted to spark dataframe. Why we didn't do in pandas and spark? Because, it was a ‘/n’ character in the title of subject column, and when it pass the new row it did not convert csv file, then pandas was tried, and some of cells values gave us the type error so it didn't use. We are making a simalarity anaylzes by using pearson correlation then, we can measure the input values of books similarities with other books.

Where: current status

- Problem : Dealing With Excel Data

```
xl_workbook = xlrd.open_workbook("lib-statistics.xlsx")
#sheet_names = xl_workbook.sheet_names()
#print('Sheet Names', sheet_names)
#xl_sheet = xl_workbook.sheet_by_name(sheet_names[0])
xl_sheet = xl_workbook.sheet_by_index(0)
#print ('Sheet name: %s' % xl_sheet.name)
data = []
for row in range(xl_sheet.nrows):
    line = []
    for col in range(xl_sheet.ncols):
        line.append(str(xl_sheet.cell(row, col).value))
    data.append(line)
#print((data))
```

```
columnNames = data[0]
dataValues = data[1:]
dataPairs = []
for row in range(len(dataValues)):
    line = {}
    for col in range(len(columnNames)):
        line[columnNames[col]] = dataValues[row][col]
    dataPairs.append(Row(**line))
#dataPairs
```

```
dataRDD = sc.parallelize(dataPairs)
df = dataRDD.toDF()
```

```
df = df.select(df['Sınıflama'], df['Eser Adı'], df['Yazar'], df['Dil'], df['Konu Başlıkları'], df['Ödünç Sayısı'].cast('float'))
df = df.fillna({'Ödünç Sayısı' : 0.0})
```

Where: current status

İSTANBUL ŞEHİR ÜNİVERSİTY

- Our Data :

```
df.show()
```

```
+-----------------+--------------------+--------------------+---+--------------------+------------+
|        Sınıflama|            Eser Adı|               Yazar|Dil|     Konu Başlıkları|Ödünç Sayısı|
+-----------------+--------------------+--------------------+---+--------------------+------------+
|      AM 7/.M8713|Museum frictions ...|                    |eng|Museums--Social A...|         0.0|
|    B 105 .I49/T34|Tahayyül gücünü y...|                    |tur|Imagination (Phil...|         0.0|
|   B 3279 .H48/D36|Bir yol var : Min...|        Damcı, Taner|tur|Ontology
Stress M...|         0.0|
|   B 415 .A9/A7519|Psikoloji şerhi =...|            İbn Rüşd|tur|Aristotle
Aristot...|         2.0|
|  B 5074 .S544/K39|An examination of...|      Kaya, Vefa Can|eng|Özel, İsmet, 1944...|         0.0|
|  B 5074 .S544/K39|An examination of...|      Kaya, Vefa Can|eng|Özel, İsmet, 1944...|         0.0|
|  B 753 .I53/R5320|Ruhun uyanışı, ya...|İbn Ṭufeyl, Muḥam...|tur|Philosophy, Islam...|         0.0|
|    B 790 .R87/C34|Çağımızın sorunla...|   Russell, Bertrand|tur|Philosophy, Moder...|         1.0|
|   BD 336 .H47/S53|Heidegger'in kulü...|         Sharr, Adam|tur|Entity (Philosoph...|         1.0|
|   BD 450 .N37/I57|İnsan ve tabiat =...|Nasr, Seyyid Hüseyin|tur|Philosophical Ant...|         0.0|
|   BF 173 .D63/E75|Erich Fromm'un ve...|     Dobrenkov, V.İ.|tur|Psychoanalysis
Fr...|         0.0|
|   BF 575 .H27/Y88|Hayatı kolaylaştı...|        Yüter, Ahmet|tur|Happiness
Persona...|         0.0|
|   BF 575 .K56/U98|Üzüntüden kurtulm...|ebu yusuf s-Sabba...|tur|Worry
Success
Üzü...|         0.0|
|   BF 635 .G43/C34|Çağdaş yaşam ve n...|       Geçtan, Engin|tur|Psychology, Patho...|         0.0|
|  BF 637 .S4/H3819|Düşünce gücüyle t...|      Hay, Louise L.|tur|Self-Actualizatio...|         0.0|
| BF 692.2 .S65/B36|Bana bilgiçlik ta...|     Solnit, Rebecca|tur|Sex Differences (...|         2.0|
|      BF 698/.F746|Sahip olmak ya da...|        Fromm, Erich|tur|Personality
Ontol...|         1.0|
|   BF 723 .P25/C83|Geliştiren anne baba|    Cüceloğlu, Doğan|tur|Parent-Child Rela...|         2.0|
|      BF 76.5/.B67|Research design a...| Bordens, Kenneth S.|eng|Psychology--Resea...|         4.0|
|  BJ 1291 .I43/I85|İslam büyüklerini...|         İmam Şa'rani|tur|Islamic Ethics
İs...|         0.0|
+-----------------+--------------------+--------------------+---+--------------------+------------+
only showing top 20 rows
```

Where: current status

- About Column of «Sınıflama»(Classification)

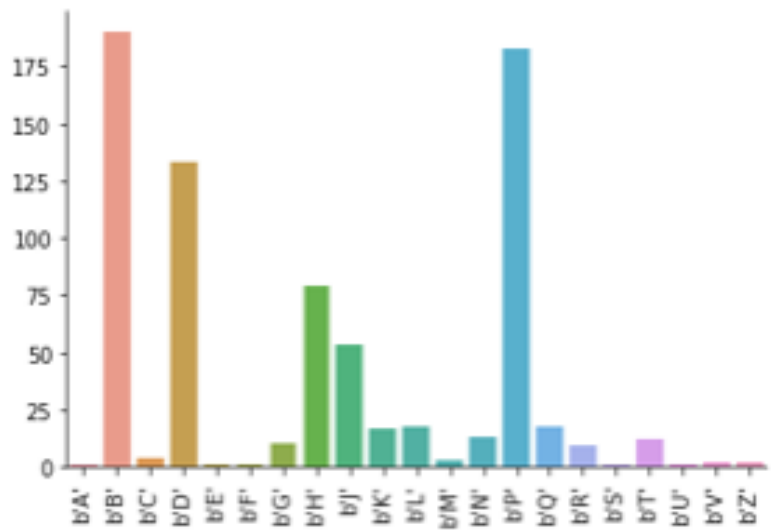
```
A -- GENERAL WORKS
B -- PHILOSOPHY. PSYCHOLOGY. RELIGION
C -- AUXILIARY SCIENCES OF HISTORY
D -- WORLD HISTORY AND HISTORY OF EUROPE, ASIA, AFRICA, AUSTRALIA, NEW ZEALAND, ETC.
E -- HISTORY OF THE AMERICAS
F -- HISTORY OF THE AMERICAS
G -- GEOGRAPHY. ANTHROPOLOGY
H -- SOCIAL SCIENCES
J -- POLITICAL SCIENCE
K -- LAW
L -- EDUCATION
M -- MUSIC AND BOOKS ON MUSIC
N -- FINE ARTS
P -- LANGUAGE AND LITERATURE
Q -- SCIENCE
R -- MEDICINE
S -- AGRICULTURE
T -- TECHNOLOGY
U -- MILITARY SCIENCE
V -- NAVAL SCIENCE
Z -- BIBLIOGRAPHY. LIBRARY SCIENCE. INFORMATION RESOURCES
```

Where: current status

- About Column of «Sınıflama»(Classification)

```
df3 = df.groupBy(split(split(df['Sınıflama'], " ")[0], "")[0].alias("group")).count().orderBy("group")
group_count = df3.count()
```

```
groups = np.empty(group_count, dtype="S30")
counts = np.empty(group_count)
for (index,row) in enumerate(df3.collect()):
    groups[index] = row['group']
    counts[index] = row['count']
sns.barplot(x=groups, y=counts)
plt.xticks(rotation='vertical')
sns.despine()
```



Where: current status

İSTANBUL ŞEHİR ÜNİVERSİTY

- Problem of The Data : Dealing with Null Values

```
def countEmptyAndNull(df):
    for i in df.columns:
        print("Column '%s' has %d empty, %d not empty, and %d null rows." % (i, df.filter(df[i]=='').count(),
                                                                              df.filter(df[i]!='').count(),
                                                                              df.filter(df[i].isNull()).count()))
```

```
countEmptyAndNull(df)
```

```
Column 'Sınıflama' has 0 empty, 751 not empty, and 0 null rows.
Column 'Eser Adı' has 0 empty, 751 not empty, and 0 null rows.
Column 'Yazar' has 294 empty, 457 not empty, and 0 null rows.
Column 'Dil' has 0 empty, 751 not empty, and 0 null rows.
Column 'Konu Başlıkları' has 1 empty, 750 not empty, and 0 null rows.
Column 'Ödünç Sayısı' has 0 empty, 0 not empty, and 598 null rows.
```

Where: current status

- Describe of The Languages

```
df.groupBy('Dil').count().show()
```

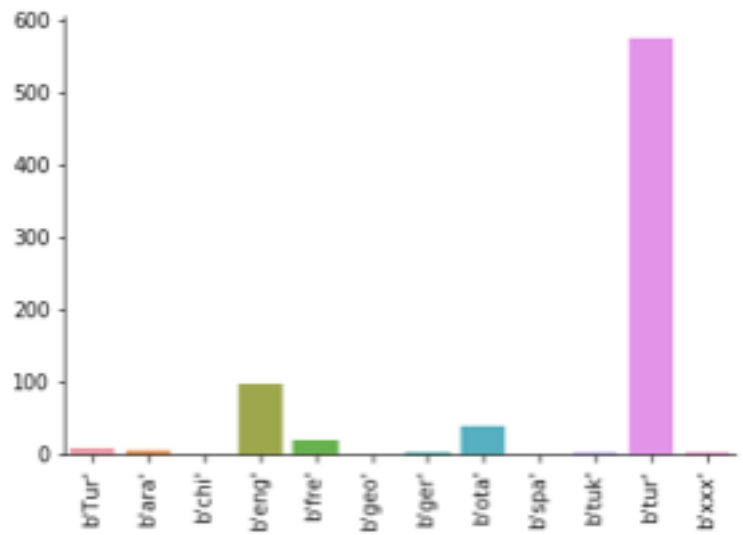
```
+---+-----+
|Dil|count|
+---+-----+
|fre|   18|
|tuk|    2|
|geo|    1|
|Tur|    8|
|ara|    4|
|eng|   98|
|xxx|    2|
|tur|  575|
|ota|   38|
|spa|    1|
|chi|    1|
|ger|    3|
+---+-----+
```

Where: current status

- Describe of The Languages

```
df4 = df.groupBy(df['Dil']).count().orderBy('Dil')
lang_count = df4.count()
```

```
langs = np.empty(lang_count, dtype="S30")
counts = np.empty(lang_count)
for (index,row) in enumerate(df4.collect()):
    langs[index] = row['Dil']
    counts[index] = row['count']
sns.barplot(x=langs, y=counts)
plt.xticks(rotation='vertical')
sns.despine()
```



Roadmap

İSTANBUL ŞEHİR ÜNİVERSİTY

- Code to cluster data for input book names according to their Classification, language, and Subject Topics
- Plot according to Pearson correlation coefficient